

CLASSIFICAÇÃO HIERÁRQUICA DE CONSENSO

Anabela Cardoso Marques

Escola Superior de Tecnologia do Barreiro
Instituto Politécnico de Setúbal

RESUMO

Neste trabalho apresenta-se a técnica Classificação segundo uma direcção, introduzida por Vichi, em 1995. Esta técnica tem por base a teoria da Análise Classificatória e dos Métodos de Consenso.

Aplicando a Classificação segundo uma direcção a um conjunto de dados tridimensionais obtém-se uma classificação hierárquica única.

Palavras-chave: Métodos de Consenso; Análise Classificatória; Análise de Dados Tridimensionais; Classificação Hierárquica

1. INTRODUÇÃO

Muitas são as instituições que recolhem informação estatística sazonalmente, por vezes a análise estatística não vai além da estatística descritiva elementar, no entanto, quando os objectivos são mais ambiciosos e se pretende conhecer o grau de semelhança das unidades estatísticas ao longo de uma série de ocorrências, verifica-se que a conhecida Análise Classificatória não pode ser aplicada a um conjunto de dados tridimensionais.

A Classificação segundo uma direcção, introduzida por Vichi, em 1995, é uma técnica para a análise de um conjunto de dados multivariados específico, que fornece ao investigador uma classificação hierárquica única para o conjunto de elementos a classificar, podendo ser considerada como sendo uma classificação síntese.

2. CLASSIFICAÇÃO SEGUNDO UMA DIRECÇÃO

Seja X um conjunto de dados multivariados definido como sendo o resultado da observação de k variáveis para n indivíduos em r ocorrências diferentes. Estes dados constituem as observações de um vector e podem ser representados num paralelepípedo, definido pelos três níveis segundo os quais os dados forem classificados: unidades estatísticas, variáveis e ocorrências.

Simbolicamente, $X = \{x_{ijh} : i \in I, j \in J, h \in H\}$, onde x_{ijh} é o valor observado da j -ésima variável para a i -ésima unidade estatística na h -ésima ocorrência, sendo $I = \{1, \dots, n\}$, $J = \{1, \dots, c\}$ e $H = \{1, \dots, r\}$.

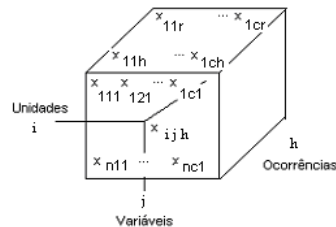


Fig. 1. Conjunto de dados tridimensionais

Os métodos descritos na Análise Classificatória não podem ser aplicados directamente ao conjunto de dados X , dada a sua estrutura tridimensional. Além disso, perante um paralelepípedo, com $n \times c \times r$ valores reais, por vezes, o investigador depara-se com o problema da elevada dimensão dos dados a analisar, levando-o a procurar técnicas para a redução dessa dimensão.

Vichi, em 1995, introduziu a Classificação segundo uma direcção, esta técnica permite fixar um dos níveis pelo qual os dados estão caracterizados, obtendo-se assim, um conjunto de L matrizes sobre o conjunto X .

No caso de fixar o nível das:

- ocorrências: \underline{r} matrizes unidades-variáveis $X_{..1}, X_{..2}, \dots, X_{..r}$ -fig. 2(a);
- variáveis: \underline{c} matrizes unidades-ocorrências $X_{.1}, X_{.2}, \dots, X_{.c}$ -fig. 2(b);
- unidades: \underline{n} matrizes ocorrências-variáveis $X_{1..}, X_{2..}, \dots, X_{n..}$ -fig. 2(c)

como podemos observar na figura 2.

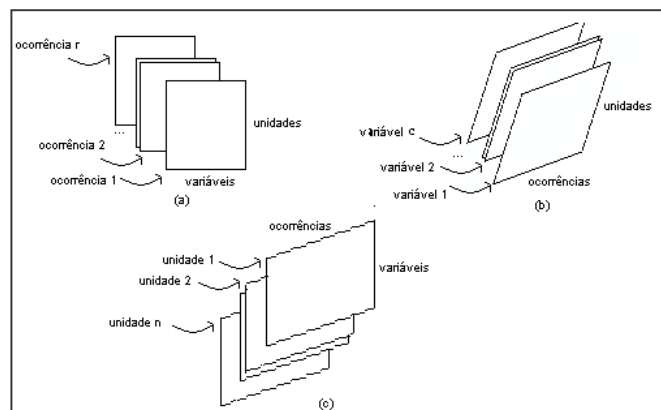


Figura 2

A escolha do nível a fixar no conjunto de dados X , depende unicamente do investigador e dos objectivos que este pretende atingir.

Com o objectivo de apresentar esta técnica, considere-se a situação da classificação hierárquica das unidades quando fixado o nível das ocorrências, uma vez que esta classificação parece ser a mais interessante do ponto de vista estatístico.

Suponhamos que estamos perante um conjunto de r matrizes unidades-variáveis para analisar. Vichi, propõem que se aplica a cada uma das r matrizes, a Análise Classificatória, obtendo-se assim um conjunto de r classificações hierárquicas, associadas a estas r classificações, vamos ter:

1) Um conjunto de r árvores binárias, T_1, T_2, \dots, T_r , onde, cada T_h é uma colecção de no máximo $2n-1$ subconjuntos de I (Figura 3), tais que:

- $\{i\} \in T_h, i \in I, i=1, \dots, n$
- $\emptyset \notin T_h$
- $I \in T_h$
- $J, J' \in T_h \Rightarrow J \subset J' \text{ ou } J' \subset J \text{ ou } J' \cap J = \emptyset$;

2) Um conjunto de r dendrogramas, $\Delta_1, \Delta_2, \dots, \Delta_r$, onde:

$$\Delta_h = \{\delta(I_{1,h}), \delta(I_{2,h}), \dots, \delta(I_{2n-1,h})\}$$

$I_{j,h}$ representa a classe obtida após a j -ésima fusão (convém não esquecer que inicialmente existem n classes formadas pelas n unidades estatísticas) e $\delta(I_{j,h})$ é o correspondente valor da fusão. Além disso, temos que:

- $\delta(I_{1,h}) \leq \delta(I_{2,h}) \leq \dots \leq \delta(I_{2n-1,h})$
- $I_{j,h} \subseteq I_{t,h} \Rightarrow \delta(I_{j,h}) \leq \delta(I_{t,h})$

3) Um conjunto de r matrizes ultramétricas, U_1, U_2, \dots, U_r , onde,

$$U_h = \{u_{ijh} : i, j \in I\}$$

e cujo elemento u_{ijh} satisfaz a desigualdade ultramétrica

$$u_{ijh} \leq \max(u_{ilh}, u_{jlh}) \quad \forall (i, l, j) \in I, (h=1, \dots, r)$$

Para encontrar a classificação hierárquica única, mais próxima das r matrizes ultramétricas, pode-se usar uma aproximação por optimização, na qual se determina a matriz ultramétrica cuja distância às r matrizes ultramétricas é mínima.

O problema em estudo pode ser descrito como sendo um problema de optimização quadrática, onde se pretende minimizar a seguinte função:

$$L(U, \lambda) = \sum_{h=1}^r \sum_{i=1}^n \sum_{j=1}^n (u_{ijh} - u_{ij})^2 + \lambda \left(\sum_{\phi} (u_{im} - u_{mj})^2 \right)$$

$$\text{com, } \phi = \{(i, j, m) : i, j, m \in I; i \neq j \neq m, u_{ij} \leq \min(u_{jm} - u_{im})\}$$

Para a minimização da função $L(U, \lambda)$, podemos utilizar um dos algoritmos de Programação Sequencial Quadrática:

- *Conjugate Gradiente*
- *Método Quasi-Newton*
- *Método Truncated-Newton*

Dada a complexidade deste problema, e o crescimento exponencial do tempo de processamento para a obtenção da solução exacta, faz sentido procurar procedimentos heurísticos mais rápidos.

Neste sentido, Vichi em 1996, com o objectivo de minimizar a função objectivo:

$$F.O = \sum_{h=1}^r \|U_h - U\|^2$$

apresenta duas heurísticas, que irão ser designados por Algoritmo A e B, respectivamente.

Algoritmo A

- Passo 1: Obter um conjunto de partições $\{ I_1, I_2, \dots, I_{r-1} \}$, aplicando à matriz que resulta da soma ponderada das r matrizes ultramétricas a Análise Classificatória.
- Passo 2: Ordenar o conjunto de partições pelo valor de fusão e aplicar a cada partição a operação “SWAP”
- Passo 3: Calcular os novos valores de fusão e escolher de entre as sub-árvores possíveis, a sub-árvore que tem o menor valor de aumento na F.O.
- Passo 4: Repetir o processo $r-1$ vezes.

Algoritmo B

- Passo 1: Para cada matriz ultramétrica ou de dissemelhança, encontrar o valor mínimo, calcular a média das dissemelhanças para as classes onde foi encontrado o valor mínimo e escolher a média mais pequena.
As classes que deram origem a esse valor irão fazer parte da nova hierarquia de partições.
- Passo 2: Calcular o valor do erro que se está a cometer ao fundir as duas classes; Actualizar as matrizes ultramétricas/matriz de dissemelhanças.
- Passo 3: Repete-se o processo $r-1$ vezes.

3. CONCLUSÕES

Neste trabalho aborda-se o problema com que alguns investigadores se deparam quando pretendem efectuar uma Análise Classificatória a um conjunto de dados caracterizados em três níveis diferentes (unidades estatísticas; variáveis e ocorrências).

Apresenta-se uma técnica que permite reduzir a dimensão dos dados a analisar e que fornece ao investigador uma classificação hierárquica única, mediante a resolução de um problema de optimização quadrática. Dada a complexidade dos problemas de programação sequencial quadrática apresenta-se dois algoritmos que permitem obter a desejada classificação de uma forma simples e eficaz.

4. REFERENCIAS

VICHI, M. (1995) – “The classification of a three-Way data set”. *Proc.Internal. Statistical Institute*, Beijing.

VICHI, M. (1996) – “Consensus of Hierarchical Classifications”, *Springer*